Institute for Research on Poverty
Discussion Paper no. 999-93

Prying the Lid from the Black Box: Plotting Evaluation Strategy for Welfare Employment and Training Programs

David Greenberg
Department of Economics
University of Maryland-Baltimore County
and Institute for Research on Poverty
University of Wisconsin-Madison

Robert H. Meyer
Harris Graduate School of Public Policy Studies
University of Chicago
and Institute for Research on Poverty
University of Wisconsin-Madison

Michael Wiseman
La Follette Institute of Public Affairs
Institute for Research on Poverty
University of Wisconsin-Madison

March 1993

The research reported in this paper was supported by the Institute for Research on Poverty and the Robert M. La Follette Institute of Public Affairs, both at the University of Wisconsin-Madison. Michael Wiseman also acknowledges continuing support from the Centre for Social Policy at the University of Bremen, Germany. The authors have benefited substantially from comments on an early draft made by Chuck Phelps and the generous provision of data by Fred Doolittle and Daniel Friedlander of the Manpower Demonstration Research Corporation and Larry Orr and Winston Lin of Abt Associates, Inc.

Abstract

To date, most evaluations of welfare-related employment and training (WRET) programs have focused on the difference between outcomes in a single site for people who receive the program's "treatment" and those in a control group who do not. This paper argues that progress in determining what makes programs effective requires greater emphasis on planning for synthesis of results from multiple sites and perhaps multiple treatment variations to provide insight into the program's *production function*: the relationship among outcomes and the characteristics of the program tested, the environment of implementation, and the people who participated. The authors develop a multilevel model of WRET outcomes and use the model (1) to determine analytically the number of sites and observations per site required to achieve a target level of efficiency in estimating production function parameters, (2) to structure a review of recent multisite WRET program evaluations, and (3) as a foundation for a suggested multisite program evaluation agenda.

Prying the Lid from the Black Box: Plotting Evaluation Strategy for Welfare Employment and Training Programs

Table of Contents

Prying the Lid from the Black Box: Plotting Evaluation Strategy for Welfare Employment and Training Programs

1. THE QUESTION

The number of evaluations of welfare-related employment and training (WRET) programs grows without sign of abatement (Blank, 1994; Greenberg and Wiseman, 1992a and 1992b; Gueron and Pauly, 1991). For the most part, these evaluations are concerned principally with the "grand mean effects" of particular innovations in particular locations, that is, with the difference between average outcomes for people who receive the program's "treatment" and those in a control group who do not. Taken individually, such studies provide little or no information on how an innovation actually produces these mean effects or how the effects might be expected to change with slight variation in program, participant characteristics, or the environment of implementation. To use a somewhat shopworn characterization, the production function for such interventions is treated as a "black box."

Given the accumulation of demonstrations, the problem of drawing inferences about the underlying production relationships could be treated as a problem in research *synthesis*. But *ex post* attempts at synthesis are inherently dependent on what serendipity has provided with respect to the innovations studied and the contexts in which they have been evaluated. The central proposition of this paper is that progress in determining what makes programs effective requires placing greater emphasis upon planning for synthesis *ex ante* and deliberately engineering the necessary variation in technique and environment to learn what works, in what circumstances, and with which types of participants.

While in some cases it is possible to experiment with different combinations of technique and participants at a single location, for administrative and other reasons it is often necessary to utilize multiple sites for this purpose. To discover the effects of variation in the larger economic and social environment, it is generally essential that multiple sites be employed. Despite its utility, site multiplication in experimental design can be expensive in both tangible and intangible ways. Obviously, costs rise with sites because of the overhead associated with setting up data collection procedures in each site. Less obvious, at least to the

outsider, are the costs of the entrepreneurial effort needed to recruit sites in a world of decentralized program administration and funding. In many cases it is a major feat of diplomacy to induce the participation of relevant agencies in each new location and to assure that methods to be tested are properly implemented.

In the face of such costs, enhancing the power of tests of grand mean effects by multiplying observations at particular sites becomes quite attractive. To justify the financial and diplomatic effort required to multiply sites instead requires careful attention to the gains to be had from such efforts. In this paper, we develop a multilevel evaluation model that illustrates the relationship between site effects and production function estimation. We use the model to identify important areas for technical improvement in the evaluation of employment and training programs, to characterize three strands of the current literature on program effects, and to frame an agenda for future research.

Much of the analysis that follows is applicable to the evaluation of employment-related training programs in general and, indeed, to any social intervention introduced in multiple locations and evaluated using a classical experimental design. However, welfare-to-work programs exhibit much greater procedural uniformity than is characteristic of other types of social interventions, and for several reasons data sources and collection procedures are often quite similar across projects. This provides opportunities often unavailable in the context of other efforts at research synthesis in social policy.

Our discussion is organized as follows. In the next section, we set out a simple model of what evaluations are about and how data from multiple sites can effectively be used. In section 3, we apply the model to study the determinants of the optimum number of sites and participants per site in studies of the determinants of program effects. The apparatus is then utilized in section 4 to characterize three strands of the literature on program effects. We illustrate each strand with an example from recent program evaluation reports. In section 5, we sketch the agenda for future policy and try to identify the next steps to which our title alludes. Section 6 concludes.

2. AN INTRODUCTION TO RESEARCH SYNTHESIS: A SIMPLE MULTILEVEL EVALUATION MODEL

Multilevel Analysis

We begin by presenting a simple multilevel statistical model that permits formal evaluation of the determinants of program net impacts. The model is referred to as *multilevel* (or *hierarchical*) because it is based on both individual-level and site-level data from multiple sites. Multilevel models have been used extensively in the education literature (see, for example, Bryk and Raudenbush, 1992, and the citations therein). As will be demonstrated below, the model is a rudimentary extension of the evaluation framework commonly used to study employment and training programs.

We set the stage by assuming that some training innovation is introduced for some class of people in several different sites--for example, in several different local welfare offices--that are sufficiently separated to assure that no significant spillover of program effects from one location to the next will occur. Information is collected on the characteristics of the clients, the environment (e.g., the unemployment rate, the skills required for local jobs, etc.), the innovation, and the outcomes of interest. We assume that the site evaluations are planned from the beginning with multilevel analysis in mind. As a result, the data collection methods used are uniform across sites. However, the methods we discuss can be applied, with some modification, to data obtained from multiple independent studies in which some variation in outcome and input measures and methods occurs. In this context, however, it is generally necessary to fall back upon less rigorous techniques of research synthesis such as narrative summary or, alternatively, what is termed *meta-analysis* (Hedges and Olkin, 1985; Hunter and Schmidt, 1990; Rosenthal, 1991; Cook et al., 1992).

As is customary in the literature on multilevel modeling, we describe the production function for the innovation with two sets of equations. The foundation is a set of *micro* models, based solely on individual-level data, for each outcome. The second set consists of one or more *macro* models, based solely on site-

level data, which relate program effects evident in the micro analysis to program characteristics and pertinent features of the site environment.

The Micro Model

The micro model is essentially identical to the evaluation model that has been used in numerous studies of welfare-to-work programs to estimate program impacts at the site level (see the descriptions in Greenberg and Wiseman, 1992a). It is given by[1]

$$Y_{ij} = \beta_j X_{ij} + \theta_j P_{ij} + e_{ij} \quad (1)$$

where i and j index individuals and sites, respectively; Y represents a programmatic outcome such as earnings or welfare payments; X represents a vector of measured individual characteristics (including a constant term) assumed to determine the individual outcome; P_{ij} is a program participation indicator equal to 1 if an individual participates in the WRET program (the treatment group) and zero otherwise (the control group); e_{ij} is an error term; β_j is a site-specific vector of parameters that captures the relationship between individual characteristics and outcomes; and θ_j is a parameter that represents the impact of the program in site j.

The model represented by equation (1) can be extended to allow the impact of a program to vary among different subgroups in the population or as a function of individual characteristics. The extended model is given by

$$Y_{ij} = \beta_j X_{ij} + (\theta_{1j} + \theta_{2j} Z_{ij}) P_{ij} + e_{ij} \ , \quad (2)$$

where Z_{ij} represents a vector of mutually exclusive subgroup indicators or a vector of individual characteristics (excluding a constant term); θ_{1j} is a parameter that represents the impact of the program in site *j* for a benchmark group; and θ_{2j} is a vector of parameters that captures the degree to which the impact of the program in site *j* differs among subgroups or as a function of individual characteristics.

If the experimental design is such that individuals at each site are assigned at random to the treatment and control groups, the impact estimates in model (1) are given simply by the average difference in outcome between the treatment and control groups in each site.[2] Impact estimates by subgroup are similarly defined. The overall impact of the program, if there are no subgroup effects, is given by the global mean $\dot{\theta} = \Sigma n_j\theta_j/\Sigma n_j$, where n_j is the total number of program participants and controls in site *j*. Global impact estimates by subgroup are similarly defined.

The Macro Model

Compared to the general literature on WRET effects, the new wrinkle in our analysis is the macro model.[3] In the context of a model with no subgroup effects, a simple macro equation is given by

$$\theta_j = \gamma F_j + w_j \ , \tag{3}$$

where θ_j is the impact parameter from the micro model; F_j represents a vector of program components or inputs, community characteristics, and economic conditions (including a constant term) assumed to affect the impact of a program in a given site; w_j is an error term; and γ represents a parameter vector. Similar macro equations are required for subgroup effects, if such impacts vary from site to site.[4] For simplicity, we

concentrate on the single macro equation given by (3). It is from estimates of the parameters of the macro equation that insight into the production function for program effects will be gained.

In this paper we do not directly address the precision of measurement of the program components F_j. However, as is discussed in Greenberg and Wiseman (1992b), very little attention is devoted in most WRET program evaluations to calibration of process or characterization of the nature of inputs grouped together under general terms like "training" or "job search." As a result, F_j is in a sense measured with error. In what follows, this problem is treated as if it simply increases w_j, and we ignore consequences of such errors for bias in the estimates of γ.

The Two Steps in Macro Model Estimation. The macro model can be estimated in one of two equivalent ways. First, it can be substituted into the micro equation (thereby eliminating the impact parameters θ_j) and estimated jointly with the remaining micro parameters and variance components. Second, the macro model can be estimated after the impact parameters $\hat{\theta}_j$ have been estimated. In this case, (3) needs to be rewritten to accommodate the fact that $\hat{\theta}_j$ is estimated with error. Thus,

$$\hat{\theta}_j = \gamma F_j + w_j + \varepsilon_j \,, \tag{4}$$

where ε_j is the error in estimating θ_j.

The usefulness of the macro equation depends upon two factors: (1) whether there is any genuine variation in program impacts across sites, and (2) whether this variation, if it exists, is related to identifiable variation in program characteristics and/or local conditions. It makes sense to address the first question, which we consider the first step in macro model estimation, before proceeding to the second, conditional, step of attempting to model the determinants of whatever significant variation exists in site effects.

Testing for Effect Homogeneity. It is common for policy analysts to confront a collection of research studies on the impact of some program by attempting to relate observed variation in calculated effects to reported differences in program features. This "narrative synthesis" approach amounts to informal estimation of the parameters γ, but in most narrative synthesis efforts there are far more program features than there are sites. The result, we suspect, is that the outcomes may be "explained" in intriguing ways even when, in fact, the variation in the estimated $\hat{\theta}_j$'s cannot confidently be attributed to anything more than noise--there is no systematic variation to explain. An important implication of this possibility is that it is potentially quite misleading to investigate informally the relationships between program characteristics and program impacts without first evaluating the case for treating intersite differences as nonrandom.

In principle, the null hypothesis that program impacts are identical in all sites can be tested by a simple Chi-square test, as is illustrated later in this paper (Rosenthal and Rubin, 1982; Hedges, 1984; Bryk and Raudenbush, 1992). As stressed in the meta-analysis literature, it is very important to conduct such a test prior to assessing the relationship between program impacts and program characteristics or community characteristics. In most evaluation contexts, it is important to know when observed differences in site effects are significant, even if explanations for such variation, if it is shown to exist, are hard to find.

Our multilevel model allows us to restate our proposition. It is our contention that the second step in macro modeling of intervention effects, that of estimating the macro parameters γ, should now become the primary target of evaluation of welfare-related employment and training programs.[5] One of the major implications of multilevel analysis (and the nested methodology of meta-analysis) is that it is possible in principle to estimate the macro parameters γ with great precision even if it is not possible, due to inadequate sample sizes at each site, to estimate the individual impact parameters (θ_j) precisely. But, as is intuitively obvious, realizing precision in estimation of the macro parameters requires examination of the association of impact variation with variation in the components of F_j. Such variation is in general achievable only with

multiplication of sites. In the next section we apply our model to evaluation of the trade-off between sites and observations per site encountered in estimation of the parameters of the macro model.

3. INDIVIDUALS VERSUS SITES IN THE DESIGN OF EXPERIMENTS

Precision

To understand the gains from site multiplication, consider the formulas for the precision of an estimated impact parameter ($\hat{\theta}_j$) and an estimated macro parameter, say, the first element of the vector $\hat{\gamma}$--call this parameter $\hat{\gamma}_1$. The variance of $\hat{\theta}_j$ is given by[6]

$$\sigma_j^2 \equiv var(\hat{\theta}_j) = \frac{\sigma_{e_j}^2 / s_{P_j}^2}{(n_j - K_1)(1 - \overline{R}_j^2)}, \tag{5}$$

where $\sigma_{e_j}^2$ is the variance of the individual error e_{ij} in site j, n_j is again the total number of program participants and controls in site j, K_1 is the number of regressors in the micro model (excluding the constant term), $s_{P_j}^2$ is the variance of the participation indicator in site j,[7] and $\overline{R}_j^2$ is the variance explained by a regression of the participation indicator on the other variables in the micro model (the vector X), as measured by the corrected R^2 statistic. Thus $\overline{R}_j^2$ measures the degree of multicollinearity between the participation indicator and X. It is zero if individuals are assigned randomly to the participant and control groups.

The formula for the precision of a macro parameter is in general substantially more complicated than in the case of an impact estimate (Bryk and Raudenbush, 1992). To simplify the discussion, we therefore assume that the number of individuals at each site is identical (equal to n); the variance of the individual error is identical in all sites (equal to σ_e^2); the fraction of individuals assigned to the control group at each site is

identical, hence, the variance of the participation indicator is identical in all sites (equal to s_P^2); and individuals are assigned randomly to the participant and control groups, hence $\overline{R}_j^2 = 0$.

Under these circumstances, the variance of $\hat{\theta}_j$ is identical in all sites (see (5)), and the formula for the variance of the macro parameter $\hat{\gamma}_1$ is given by

$$\sigma_1^2 \equiv var(\hat{\gamma}_1) = \frac{\dfrac{\sigma_W^2}{(J-K_2)} + \dfrac{\sigma_e^2/s_P^2}{(J-K_2)(n-K_1)}}{s_{F1}^2(1 - \overline{R}_1^2)}, \qquad (6)$$

where J is the total number of sites, K_2 is the number of regressors in the macro model (excluding the constant), σ_W^2 is the variance of the error in the macro equation, s_{F1}^2 is the variance of the regressor corresponding to the macro parameter γ_1 (the first variable in the vector *F*), and $\overline{R}_1^2$ is the variance explained by a regression of the regressor F_1 on the other variables in the macro model, as measured by the corrected R^2 statistic. This statistic captures the level of multicollinearity among the variables in the macro model. The term to the left of the plus sign in (6) reflects the error *w*--the part of θ_j that is not explained by the observed program and community characteristics. The term to the right of the plus sign in (6) reflects the error ε--the error in estimating $\hat{\theta}_j$. Formulas for the variance of other estimated macro parameters are similarly defined.

Optimization

Given estimates of the parameters in (6) and information on the relative costs of collecting data at given versus additional sites, it is possible to compute the number of sites (J^*) and the number of individuals per site (n^*) that are required to obtain a target level of precision (σ_1^2) at minimum cost.[8] As a first step in solving for the optimum values of *J* and *n*, note that (6) implies that the optimum number of sites is given by

$$J^* = J_1^* + J_2(n^*) \ , \tag{7}$$

where

$$J_1^* = \frac{1}{\sigma_1^2} \cdot \frac{\sigma_w^2}{s_{F1}^2} \cdot \frac{1}{(1 - \bar{R}_1^2)} + K_2 \tag{8}$$

and

$$J_2(n^*) = \frac{\sigma_e^2/s_P^2}{(n^* - K_1)\sigma_1^2 s_{F1}^2 (1 - \bar{R}_1^2)} \quad . \tag{9}$$

As indicated, the optimum number of sites is the sum of two terms, the first of which does not depend on n^* and thus does not depend on the relative costs of data collection for individuals and sites. We examine this term first.

Equation (8) reinforces the obvious: the number of sites (*J*) must exceed the number of regressors (i.e., program and community characteristics) included in the macro equation. If the number of regressors is greater than the number of sites, the macro parameters are simply not identified. Of course, having more sites than regressors does not guarantee identification. Equation (8) also indicates that J_1^* must be increased if any of the following three factors are increased: (1) the number of regressors in the macro model (K_2); (2) the variance (σ_w^2) of unknown (and perhaps random) determinants of program impact relative to the variance

(s^2_{F1}) of one of the known determinants of program impact; and (3) the level of multicollinearity in the macro model (with respect to the single regressor F_1). Later in the paper, we explore whether it is possible, as part of the design of an evaluation, to control these factors so as to keep J_1^* as low as possible. This is important, because, as is demonstrated below, J_1^* could be quite large under fairly common circumstances, perhaps greater than the available number of sites.

The second term, $J_2(n^*)$, depends on the unknown optimum number of individuals per site. As n^* approaches infinity, $J_2(n^*)$ approaches zero. Hence, if the number of individuals per site is quite large, the optimum number of sites is approximated by J_1^*, but this is always an underestimate.

Depending on the relative costs of collecting data at given and new sites, there is some number of sites and individuals that achieves a target level of precision at minimum cost. To address this issue, we assume that reasonable estimates of the components of (8) are available. We also assume the cost function for collecting data on sites and individuals is linear in both dimensions and is given by

$$C = k(nJ) + lJ \ , \tag{10}$$

where k is the cost of collecting data on a single additional individual at a given site and l is the fixed cost of collecting data at a new site. Since it is the relative costs of data collection that matter, it is convenient to rewrite the cost function as

$$C = k(m + n)J \ , \tag{11}$$

where $m = l/k$, the relative costs of collecting data at a new site versus collecting data on a single individual at a given site.

Substituting equations (8) and (9) into (7) and the result into the cost function yields a cost function that depends on n^*, not on J^* (see Appendix A). As indicated in Appendix A, this cost function is U-shaped. Hence it is possible to solve for the cost-minimizing value of *n*. This is given approximately by[9]

$$n^* = \left(\frac{m\sigma_e^2}{s_P^2 \sigma_w^2} \right)^{1/2} . \tag{12}$$

The interpretation of this formula is straightforward. It is cost-effective to collect data on additional individuals at a given set of sites if (1) the fixed cost of collecting data at an additional site is high relative to the cost of collecting data on additional individuals and (2) if the variability of the error in the micro model (σ_e^2) is large relative to the variability of the macro model error (σ_w^2).

This apparatus can be used to illustrate the mix of sites and observations per site required to estimate some macro parameter γ_1 with a target level of precision (σ_1^2). Given the common concern with tests of the hypothesis that the true population parameter γ_1 is zero, the target level of precision should depend on the hypothesized magnitude of γ_1. If γ_1 is quite small, it should be estimated with greater accuracy than if it is large.

What is a reasonable value of γ_1? The answer to that question cannot be determined without specifying the exact regressor associated with γ_1. Since the substantive importance of a macro parameter depends on the variance of the associated regressor (s_{F1}^2) and the variance of true program impacts across

the universe of sites (call this σ_θ^2), it is useful to focus on the beta coefficient (call this g_1) rather than the regression coefficient γ_1. The beta coefficient is defined by

$$g_1 = \frac{s_{F1}}{\sigma_\theta} \gamma_1 \ , \tag{13}$$

and the precision of the regression coefficient is given by

$$\sigma_1^2 = \frac{\sigma_\theta^2}{s_{F1}^2} \cdot \sigma_{g1}^2 \ , \tag{14}$$

where σ_{g1}^2 is the precision of the beta coefficient g_1.

Substituting (14) into (8) yields the following expression for J_1^*:

$$J_1^* = \frac{1}{\sigma_{g1}^2} \cdot \frac{\sigma_w^2}{\sigma_\theta^2} \cdot \frac{1}{(1 - \overline{R}_1^2)} + K_2. \tag{15}$$

The modified equation for $J_2(n^*)$ is:

$$J_2^* = \frac{1}{\sigma_{g1}^2} \cdot \frac{\sigma_e^2}{\sigma_\theta^2} \cdot \frac{1}{s_p^2 \ (n - K_1)(1 - \overline{R}_1^2)} + K_2. \tag{16}$$

Equation (15) recasts the formula for J_1^* in terms of the precision of g_1 and the ratio of σ_w^2 to σ_θ^2. The latter ratio is equal to one minus the explanatory power of the pure macro model--see equation (3). Equation (16) recasts the formula for J_2^* also in terms of the precision of g_1 and the ratio of σ_e^2 to σ_θ^2.

Finally, what is a reasonable value of σ_{g1}^2? Suppose that the target level of precision is set such that the *t* statistic for the beta coefficient is equal to t^* (for example, $t^* = 2.0$). This implies that the precision of the beta coefficient must equal

$$\sigma_{g1}^2 = \frac{g_1^2}{t^{*2}}. \tag{17}$$

Given alternative values of g_1, t^*, and the other parameters in (15), it is possible to determine the optimum of J_1^*. n^* and $J_2(n^*)$ can similarly be determined, as is indicated in Appendix A. Table 1 reports values of these variables for three alternative sets of assumptions about the underlying parameters. We characterize the three assumption sets in relation to their implications for the number of sites required: "unfavorable," "reasonable," and "favorable."

Consider the list in Table 1. In setting hypothetical values, we begin by assuming the target t^* of 2.0, the magic number for publication. For a "reasonable" value of g_1 we take .2, and we label .1 unfavorable and .4 favorable. These values establish the value of σ_{g1}^2. For $\sigma_w^2/\sigma_\theta^2$, the share of the variance of the impact of the intervention that is not related to variation in the regressors, we accept as a reasonable value .4 and label .6 unfavorable and .2 favorable. We expect the random component in individual outcomes to be greater than variation in intervention effects across sites; we choose a value of 2.0 for $\sigma_e^2/\sigma_\theta^2$ as a reasonable number, select 1.0 as a favorable possibility, and 4.0 as unfavorable. Fixing $\sigma_w^2/\sigma_\theta^2$ and $\sigma_e^2/\sigma_\theta^2$

establishes the value for σ_e^2/σ_w^2. To finish up, we take as a reasonable estimate of $\overline{R}_j^2$ a value of .3, a favorable value could be .1, and an unfavorable one might be .7. We assume the same values for both the micro and macro model in each case, and we assume

TABLE 1

The Optimum Number of Sites and Individuals Per Site Given Alternative Parameter Values

	1 "Unfavorable"	2 "Reasonable"	3 "Favorable"
Assumed Parameter Values			
g_1	0.1	0.2	0.4
t^*	2.0	2.0	2.0
σ^2_{g1}	0.0026	0.0104	0.0416
s^2_P	.25	.25	.25
$\sigma^2_w/\sigma^2_\theta$	0.6	0.4	0.2
$\sigma^2_e/\sigma^2_\theta$	4.0	2.0	1.0
σ^2_e/σ^2_w	6.7	5.0	5.0
$1 - \overline{R}^2_1$	0.3	0.7	0.9
K_1	5	5	5
K_2	8	8	8
m	1,000	600	100
Optimum Values			
J^*_1	776.3	62.9	13.3
$J^*_2(n^*)$	129.8	11.2	4.4
J^*	907	75	18
n^*	163	103	29
n^*J^*	147,841	7,725	522

Source: Calculated from formulas presented in text. The site requirement (J^*) is rounded upward to the next larger integer.

that random assignment is conducted in such a way that a participant stands an equal chance of being assigned to the control or treatment group.

For *m*, the ratio of the cost of establishing a site to collecting data on additional individuals in existing locations, it is difficult to select reasonable values, in part because data of this sort are rarely published and in part because it is difficult to attach dollar values to some of the political costs of site recruitment. In general the political dimension of such tasks is assumed by sponsoring units of government, so we concentrate on the hard costs of setup once such negotiations are completed. Data provided by two firms which specialize in program evaluation produced estimates of roughly 100 and 600[10]; we take the former as "favorable," the latter as reasonable, and 1,000 as an unfavorable ratio estimate. "Favorable" in this case means something a bit different from the earlier characterizations. Here a favorable assumption is associated with low costs for additional sites (holding the marginal cost of adding observations within a site constant). But relatively low additional site costs (compared to the incremental cost of observations) actually increase the solution value for sites required.

As is indicated in Table 1, given the range of parameter values selected for this exercise, the optimum number of sites ranges from 18 to an extraordinary 907.[11] Similarly, the optimum number of individuals per site varies from 29 to 163. We have so far considered site costs only relative to the incremental cost of adding an observation within a site. If this cost does not vary across assumption sets, then minimum costs may be compared across assumption sets. Calculated in this way, costs fall by a factor of approximately 20 in the move from "unfavorable" to "reasonable" assumptions, and by an additional factor of approximately 20 in the leap from "reasonable" to "favorable."

These results support our view that it is not possible to estimate reliably the parameters of a macro model without selecting a relatively large number of sites, at least twenty, and exerting careful control over experiment implementation. Confirmation of our view requires more information on the various factors that determine site requirements.

On the bright side, review of equation (8) indicates that the required number of sites can be reduced by:

(1) Accepting reduced precision. Dropping *t** from 2 to 1.68 in the examples presented would reduce the number of sites required by 23 percent in the "reasonable" case.

(2) Constraining, if possible, variation in impact associated with factors not accounted for in the macro equation (this would diminish σ_w^2). Reducing $\sigma_w^2/\sigma_\theta^2$ from .4 to .2 in the "reasonable" case would reduce the number of sites required from 75 to 46.

(3) Where possible, assigning variation in interventions randomly to sites, and minimizing adaptive responses by sites to the treatment to which they are assigned. Both of these steps would reduce multicollinearity (i.e., they would decrease $\overline{R}_1^2$). A reduction in $\overline{R}_j^2$ from .3 to .1 in the "reasonable" case in Table 1 would lower the required number of sites from 75 to 61. The problem of adaptive response is discussed in more detail later in the paper.

4. THE CURRENT LITERATURE: THREE TYPES OF OUTCOMES

Our simple model provides a way to begin categorizing the currently available stock of program evaluations by analytic perspective. We draw a distinction between evaluations oriented toward estimation of site means (θ_j in the notation developed above), evaluations oriented toward global means ($\dot{\theta}$), and evaluations aimed at uncovering the production function for program effects (γ). We conclude that, judging from the examples we have studied, the material for satisfactory hierarchical analysis has yet to be produced in employment and training evaluations, and we attempt to explain why.

Estimation of Site Means: The GAIN Evaluation

In May 1992, the Manpower Demonstration Research Corporation (MDRC) released interim findings from an evaluation of the impacts of the Greater Avenues for Independence program, a California welfare-to-work innovation initiated in 1986 that was a precursor to the national JOBS program (Riccio and Friedlander, 1992). As is a common feature of MDRC's research strategy, the GAIN evaluation was based on a classical experimental design in which program entrants were assigned at random to either a control group not provided exceptional services or a treatment group enrolled in GAIN. GAIN's impacts were measured by comparing the average first-year earnings and AFDC payments for treatment and control members in each of six counties.

At first blush, the early MDRC GAIN results seem to suggest great variation in impact on both earnings and welfare receipt. The variance in earnings impact is most substantial, ranging from a 65 percent positive difference between treatment and control groups in Riverside County to a 1 percent negative difference in Los Angeles. Much of the remainder of the report is devoted to a narrative broaching of the macro factors (F_j) possibly responsible for the intercounty differences and to investigation of the variation in program effects by participant characteristics.

The GAIN evaluation poses several difficulties. One set of problems is related to possible sources of intercounty variation. Considerable latitude was provided counties in GAIN implementation, and MDRC observers are emphatic in their claim that the programs developed in the counties studied differed substantially in many dimensions. It is natural that both the research and political communities should be concerned with the relationship between the outcomes and program structural variation and the relationship between outcomes and county economic environments. The problem with exploring these factors is that the evaluation was limited to only six counties, but relevant differences among these counties number far more than six. Indeed, *fourteen* separate explanatory variables, as well as interactions among these variables, were examined (Riccio and Friedlander, 1992, chapter 6). The authors explore these factors by presenting an

arithmetic average of county impacts and then studying outliers (Riccio and Friedlander, 1992, pp. x-xi). But given the substantial variation in impacts observed in earlier studies of welfare-to-work programs, the mere presence of variation in outcomes across sites is not *prima facie* evidence of differential effects resulting from program variation.

An alternative approach would have been to analyze the intersite variation first, using the formal methods of meta-analysis to confirm or deny the presence of variation in site effects significantly in excess of what would have been anticipated from sampling error alone. In the historical progression of MDRC methodology, the 1992 report is exceptional in that it does in fact contain the first formal evaluation of intersite differences. (The influential earlier MDRC syntheses have all been entirely narrative in character; see Gueron and Pauly, 1991.) The authors conducted pairwise comparisons of combined county impacts for AFDC-R and AFDC-UP participants using a variant of the Newman-Keuls test (Kirk, 1982, pp. 123-25) to identify exceptional differences. They report that the only statistically significant outlier was Riverside county (Riccio and Friedlander, 1992, p. 127). This test, however, is not what really is needed: what is required is a combined test of the homogeneity of observed impacts.

The meta-analysis approach involves computing[12]

$$H_T = \sum_j (v_j)(\theta_j - \bar{\theta})^2 \ , \tag{18}$$

where H_T is the weighted sum of squares of the effect size estimates θ_j about the weighted mean effect,

$$\bar{\theta} = \frac{\sum_j v_j \theta_j}{\sum_j v_j} \ . \tag{19}$$

which we have designated by $\bar{\theta}$. The weight for each θ_j, v_j, is the inverse of the variance of the estimate, and

If in fact all sites share a common treatment effect (that is, the "true" θ_j's are equal), then the statistic H_T has approximately a χ^2 distribution with J (the total number of sites) - 1 degrees of freedom. For average total earnings gains over quarters 2 through 5 following client entry across the six GAIN sites, H_T is 43.6, well above the critical (95 percent) value of the χ^2 with 5 degrees of freedom of 11.07; for reduction in AFDC payments, H_T is 29.6, also clearly significant. This provides license--already taken--for MDRC to proceed with at least "speculative estimation" of the macro equation.

In addition to analysis of site-to-site variation, the GAIN report also investigates program effects for recipient subgroups defined on the basis of performance on a literacy test and on the basis of welfare history. Subgroup effects are compared across sites, but no global mean subgroup effect is calculated. As Greenberg and Wiseman (1992b, pp. 99-102) have pointed out, it is very difficult to interpret the results of such subgroup evaluations. Lacking evidence on characteristics of the program presented to each subgroup, it is impossible to determine whether subgroup differences are attributable to variation in the productivity of inputs by participant characteristics or to variation in the inputs provided members of each subgroup.[13]

The GAIN study is an excellent illustration of both the advantages provided for efforts at synthesis by the current methodology for studying welfare-to-work interventions and the problems that arise as a result of focusing demonstration evaluation plans upon site mean effects. On the one hand, most recent studies of welfare-to-work interventions have been oriented toward impacts on earnings and public assistance receipt, and given that the institutional sources for this information are the same across states, the dependent variables tend to be comparable, especially in the MDRC work. On the other hand, program variables are measured very coarsely, typically as a dummy variable indicating treatment group assignment. Perhaps in part because of biases introduced by this error in specification, estimated mean impacts are usually small. To obtain sufficient power to achieve statistical significance at the site level for effects of the magnitude normally observed, each site must have a sizable sample. In the context of an overall budget constraint, the

consequence is that the number of sites must be small. But a small number of sites precludes reliable inferences about the effects of the program and community characteristics commonly hypothesized to affect programmatic outcomes, even if precision in measuring inputs were enhanced. Indeed, as indicated in section 2, a more formal approach to the analysis--a multilevel approach--would reveal that six sites are an insufficient number to extract information about the effects of the range of program and community characteristics cited by MDRC in narrative synthesis of the GAIN evaluation. The JOBS evaluation, which is also conducted by MDRC using basically the same methodology, is certain to present the same problems.[14]

Evaluation of the Global Mean: The National JTPA Evaluation

Title II-A of the Job Training Partnership Act of 1982 funds the nation's major training program targeted at the disadvantaged population. The National JTPA Study is an evaluation of this program (Hotz, 1992). The study, which is based on a sample of over 20,000 program applicants who were randomly assigned to either treatment or control status, estimates program impacts on both adults and out-of-school youths enrolled in JTPA in sixteen different local service areas. At this writing, findings are available for impacts on employment and earnings during the first eighteen months after acceptance into the program (Bloom et al., 1992). It is these findings that we focus upon here.

The National JTPA Study is of special interest from the perspective of this paper, because in marked contrast to the GAIN evaluation, relatively little emphasis is placed upon site-specific impacts (θ_j). Instead, the focus is upon pooling the research sample across sites and estimating global mean impacts ($\overline{\theta}$). In addition, however, considerable emphasis is placed on the analysis of program effects on various subgroups. For example, results are always reported separately for male and female adults and for male and female out-of-school youths. In addition, within each of these four groups, separate impacts are estimated for subgroups defined on the basis of ethnicity, work experience, educational achievement, training history, public assistance history, family income level, and family composition. And tests are performed to determine

whether differences in impacts among these subgroups are statistically significant. Finally, separate impacts were obtained for three groups that were categorized prior to random assignment on the basis of the training and employment services that JTPA intake staff recommended for them.

The published early findings from the JTPA study imply that the program increases the earnings of many male and female adults, has negligible effects on the earnings of out-of-school female youths, and actually causes the earnings of out-of-school male youths to decline. If these findings hold up once the full analysis is completed, they are of obvious policy importance, suggesting, for example, that new approaches need to be developed for out-of-school youths.

However, results from the National JTPA Study provide only limited information in response to the question broached by evaluation synthesis: "What works for whom?" The reasons for this are embedded in the overall evaluation design. First, the assignment of particular JTPA services to members of the treatment group was not independent of participant characteristics. The evaluation found that adults recommended for job search assistance, on-the-job training, or both tended to enjoy larger earnings improvements than those directed toward other services such as classroom training. However, as the evaluators recognize, adults recommended for particular types of services not only differed in the types of program activities for which they were recommended, but in personal characteristics as well. For example, adults recommended for on-the-job training or job search tended to be more job ready, on average, than those recommended for classroom training. Thus, one cannot be sure whether adults who received classroom training might have been more successful had they received on-the-job training instead.

Second, there are only sixteen sites. This is, of course, an appreciable number, more than that used in most random assignment evaluations of training and education programs, and perhaps too many to allow useful narrative synthesis to be conducted. However, calculation of H_T for each demographic subgroup indicates that in no case can the hypothesis of effect homogeneity across sites be rejected.[15] Therefore, there appear to be too few sites to permit meaningful analytic synthesis through estimation of macro equations

such as those described in section 2,[16] and it is possible that such effects are absent. One cannot determine whether site-specific differences in program attributes or environmental characteristics affected program outcomes in this demonstration.

Third, although an initial attempt was made to select study sites on the basis of a stratified random sample, this attempt was ultimately abandoned because of difficulties in obtaining cooperation from selected sites (Bloom, 1991, p. 112; Hotz, 1992, pp. 94-95). In fact, fewer than 20 percent of the sites initially contacted by the evaluators conducting the study ultimately agreed to participate. To enlist as many sites as possible, the JTPA evaluators have: (1) financially reimbursed sites for the costs they incur; (2) publicly recognized participating sites; (3) provided sites with technical assistance in operating JTPA; and (4) promised participating sites specific information on how well their program is "working."[17] Hence, although the sixteen study sites are diverse, they are not necessarily representative of local service areas nationally.

The experience of the JTPA study illustrates the ambiguity that is encountered in applying our site optimization model in practice. Combining the costs of the two evaluation firms (Abt and MDRC) involved plus operations payments made to participating sites produces a fixed cost estimate of approximately $239,000 per site plus a variable cost per observation of $384.[18] Thus the ratio *m* is $239,000/$384, or 622, and this is the source of the "reasonable" estimate of 600 that is used in Table 1. But these estimates reflect the cost of tangible operations, and not the cost of intangible diplomacy. Were all the resources employed to recruit sites included in this evaluation, the fixed cost would probably be greater, and it appears that the marginal cost of recruitment is rising as the pool of "unrecruited but convinceable" sites diminishes.

Macro-Level Relations: The Food Stamp E&T Program Evaluation

The Food Stamp Employment and Training (E&T) Program, which was implemented in 1987, is a mandatory program for able-bodied food stamp recipients who are under the age of sixty and not attending school or working. The greatest emphasis in E&T, by far, is on training participants in job search techniques and in monitoring participants to ensure that they fulfill a program obligation to make a stipulated number of

job contacts within a specified time period. However, a modest number of E&T participants receive education or training, and still others are assigned to work programs at government and nonprofit agencies.

The evaluation of E&T, which was initiated shortly after the program itself begun, was ambitious in scope (see Puma et al., 1990). Around 13,000 individuals who had been certified for food stamps in fifty-three separate sites in twenty-three states were randomly assigned to either treatment or control status. The fifty-three sites were selected from a stratified random sample, with the probability of selection positively related to the size of each potential site's E&T-eligible population. Separate ordinary least squares regressions were estimated for over a dozen outcome measures, five different time frames, and the fifty-three sites--a total of around 35,000 micro equations in all. Not surprisingly, given their number, these separate impact estimates were not reported. (Indeed, they were not saved and, consequently, are not available for purposes of the present study.) Instead, global means were computed by multiplying the impact estimate for each site by its sampling weight. In addition, and rather uniquely in evaluations of employment and training programs, macro equations were estimated to determine whether differences in site-level impacts were influenced by variations in site economic conditions, characteristics of the target population, and the treatment services provided. Thus, the final report contained estimates of both $\dot{\theta}_{\cdot}$ and γ, but not θ_j.

From a policy perspective, findings from the E&T impact analysis were disappointing. No evidence was found that the program increased earnings or employment among participants. Moreover, although the findings implied that the receipt of transfer payments fell, this decrease was slight. Food stamp benefits fell by about $65 during the first year after random assignment, but this appeared to be partially offset by small increases in cash assistance. One possible explanation for these findings is that program expenditures were very low, averaging only $135 per program participant. As a consequence, the E&T program increased the amount of employment and training services received by the target population, but not by very much. For example, during the year over which the evaluation was conducted, 43 percent of the treatment group

received some employment or training services, but so did 31 percent of the control group. Thus, a very ambitious and elaborate evaluation was conducted of a very modest policy intervention.

None of the coefficients in the E&T macro equations even approached statistical significance. Indeed, few coefficients in these equations exceeded their standard errors in value. In assessing this result, it would have been helpful to know the amount of variation in the site-specific impact means (σ^2_θ). It would have also been useful to know the proportion of σ^2_θ attributable to sampling error, since only the remaining variation can potentially be "explained" by the independent variables in a macro equation. This information was not provided, however, and since the values of θ_j and their standard errors are not available, it cannot be obtained.

In conducting the evaluation of the E&T program, the researchers encountered several serious operational problems that are particularly relevant to this paper. First, the E&T evaluation placed considerable emphasis on obtaining a nationally representative sample. However, despite considerable effort at site recruitment and the mandatory character of the program being evaluated, thirteen of the sixty sites originally selected for the evaluation refused to participate. In some, but not all, of these instances, similar back-up sites were randomly selected and successfully recruited to take their place.

Second, because of the large number of study sites involved in the evaluation, responsibility for random assignment in each site was given to a site coordinator, who was an employee of the local agency that administered food stamps. A majority of the sites encountered problems in implementing the random assignment procedures. In two instances, these problems were sufficiently severe that the sites were dropped from the study.

Third, to obtain outcome measures, most recent studies of employment and training programs have relied on computerized administration data--for example, welfare records and quarterly earnings reports that are required from employers for purposes of administering state Unemployment Insurance programs.

Relying on such data, however, becomes less practical as more states, and, accordingly, more administrative agencies, are involved. Separate arrangements have to be made with each agency, and computer programs have to be tailored to each set of records. Consequently, the E&T evaluation relied instead on three interviews conducted approximately every four months. Unfortunately, the interview response rate was relatively low, with all three interviews being completed for slightly less than half the research sample. Low response rates such as these will, unfortunately, bias the impact estimates if non-responding members of the treatment group differ in employment, earnings, receipt of welfare, and so forth from non-responding members of the control group.

- - -

These examples illustrate the applicability of our analysis framework, but they also reveal how primitive the research that has been conducted at the macro level is. Given that local operators are currently and are likely to continue to be called upon to make decisions about the constitution of WRET programs, work in this area seems to be the frontier.

5. PLANNING FOR SYNTHESIS *EX ANTE*: A PROPOSAL

To this point, we have concentrated on adapting the framework of hierarchical linear modeling to the context of evaluation of welfare-related employment and training programs. We use this framework to characterize the objective of WRET evaluation strategy: identifying the effects of technique, environment, and client characteristics on program outcomes. We have demonstrated that such identification ultimately will require the multiplication of sites. The bad news comes from Table 1 and from experiences with the demonstrations we have reviewed: further refinement of production function estimation in welfare-to-work experiments requires demonstration implementation at a substantial number of sites, more than are typical of most demonstrations, and multiplication of sites is not easy. But our model also shows what to look for in opportunities for next steps: other things equal, we want to focus upon innovations for which the number of

sites required for parameter estimation is financially and diplomatically feasible. The remainder of this paper is devoted to a survey of the path ahead, some observations on promising opportunities, and some caveats born from the accumulation of experience in program evaluation in the area of welfare-related employment and training programs.

The Issues

We begin by recognizing that a great deal of variation is possible in the "treatment" experienced by participants in welfare-related employment and training programs. Although, at least in superficial detail, this variation in treatment is most likely to occur across treatment sites, considerable variation is also often found within sites. Such variation has important implications for both the micro and the macro equations used in analytic synthesis and, in consequence, plotting future evaluation strategy. Below, we outline some of the dimensions along which treatment may vary.

- *Service components*

A given program participant may receive one or more of several services. Among these are directed job search, which may be provided on either an individual basis or in a group setting; community work experience, in which the participant is assigned to a government or nonprofit agency at no cost to the agency; skill-training in a classroom setting; remedial or basic education, which may emphasize an academic or more applied approach; education directed at a degree or certificate; and subsidized employment, in which private sector employers are paid a subsidy for hiring the individual. Depending upon the program, the choice of services may be made by program participants, by program managers, or by some combination of the two.

- *Delivery of services*

Program services may be delivered by government agencies, private agencies, or by some combination. When service delivery is privatized, some form of performance contracting may be used.

- *Initiation of treatment*

After entry to an employment and training program, actual receipt of services may begin immediately or be delayed for several months. This is important because the character of the population of persons who ultimately receive services will be affected. The longer the delay, the larger the proportion of the potential pool of participants who will find jobs on their own (or will leave the program for other reasons) without receiving services. And those persons who do find jobs prior to receipt of services are likely to differ systematically from those who do not.

- *Treatment length*

The hours a given service is received may vary among participants for several reasons: the number of hours the service is provided per week varies, the number of weeks individuals are allowed to receive the service differs, participant decisions as to when to leave the service (for example, as a result of finding a job) diverge, and individual attendance varies.

- *Alternative sequencing*

The order in which a set of services is received may vary. To take a simple example: an individual might first participate in job search; followed, if the individual does not find a job, by remedial education; followed, perhaps, by a second round of job search. Alternatively, remedial education could come first and then be followed by job search. Finally, the individual might instead participate in job search, perhaps for an hour or two a day, while receiving remedial education. Sites may differ in how they assign individuals to sequences. At one extreme, a single fixed sequence might be selected that is then applied to all program participants at a given site. At the other extreme, sequences might be tailored to each individual, with

decisions concerning the next service to provide depending upon the individual's performance in the preceding service.

- *Tracking strategies*

Depending on their personal characteristics, individuals may be assigned to different services, to different treatment lengths, or to different treatment sequences. For example, persons deemed "job ready" might be assigned to job search, while those judged "not job ready" might be assigned to remedial education.

- *Penalty and reward structures*

Program participants may face different penalty and reward structures. For example, in the case of mandatory programs for AFDC recipients, one welfare agency may typically reduce benefits when households refuse to participate in a program, while another may rarely invoke such sanctions, even though they nominally exist.

- *Inputs*

Even if the treatments received by two different program participants are identical along all the dimensions defined above, they may still vary in terms of programmatic inputs.[19] For example, classroom instruction might vary in terms of equipment, the number of trainees per instructor, and the relevant experience and educational background of the instructors. Or intensive case management, requiring that relatively few program participants be assigned to each caseworker, may or may not be used to guide participants toward the "right" set of services and to ensure that they actually receive the services they are assigned.

Given the different program dimensions listed above, it should be apparent that the number of possible treatment permutations is enormous. Indeed, each program participant could potentially receive a distinct "treatment." Ideally, one would like to measure the effectiveness of each possible treatment permutation. Moreover, one would also like to learn whether the relative effectiveness of the different

treatment permutations varies across different types of individuals and different local economic conditions. As a practical matter, this is obviously not feasible. Choices must be made. Thus, what we can learn from even the best multilevel evaluation is inherently limited.

The analysis in sections 2 and 3 and the experience reported in section 4 suggest that choices among the potential objects for multilevel experimental evaluation should be evaluated in light of seven considerations:

(1) The issues should be widely recognized as important among program operators, and it should be possible to explain their significance persuasively to the political establishment.

(2) There should exist reasonable grounds for believing that the productivity of the prospective WRET program is affected by variation in client, administrative agency, or environmental factors that is inadequately represented by a small number of sites and/or cannot be feasibly reproduced in a single site.

(3) It should be possible to identify the essential characteristics of the inputs involved and, to the extent that variability can occur within this input set, it should be possible to measure the variation.

(4) Implementation of the treatment to be evaluated and the data collection procedures required to measure both inputs and outputs should be known to be feasible on the basis of either prior experience or pilot operation.

(5) The likely effect of the treatment to be evaluated should be substantial, given the range of feasible variation, or, alternatively, the discovery that effects are, as suspected, modest should have substantial budgetary consequences.

(6) A means should be available to motivate participation of the WRET program operating agencies in the sites selected for utilization in the experiment and to assure that such agencies will not change the essential features of the program being evaluated.

(7) The expected value of the knowledge to be gained should be commensurate with the costs.[20]

Overview of the Proposal

Given these criteria, what might the agenda look like? We suggest consideration be given to multisite evaluation of the following treatment variations. Others may, of course, select a different agenda.

Delay. As is well known, a substantial number of welfare cases close within a short period after opening. In at least some of these cases, the consequence is a waste of resources used to evaluate needs, to organize job search or training programs, and generally to schedule activities. Most demonstration efforts in this area have emphasized the utility of early intervention; investigation of the consequence of systematic delays in treatment intervention would also seem appropriate. This intervention meets the standards set out above quite easily: the issues are widely appreciated and easy to explain; it is likely that the effect of a delay on clients is dependent upon both other site administrative features (what lies at the end of the waiting period, for example) and job market circumstances that can be studied most readily in the context of site-to-site variation; the intervention is readily measured; and if it has no consequences for overall productivity, substantial cost savings may be involved. Here is a case, however, in which what we have termed the problem of "adaptive response" comes into play. If indeed a required delay saves resources, then most agencies can be expected to use the resources to alter the treatment that follows the waiting period. While undoubtedly considerable variation from site to site in the post-delay activities will still exist, even with control for what goes on after, this response will reduce the efficiency of estimates of the delay effect. One strategy for suppressing such responses would be to insist that money saved be directed to programs for groups not targeted by the intervention--for example, long-term recipients.

Tracking. A major focus of the current JOBS evaluation is comparative study of the effects of two approaches to recipient tracking. One emphasizes early experience in job search, followed by more substantial training programs for clients who fail to find employment. The alternative is to begin the WRET program with a more comprehensive evaluation of client needs and to route those with substantial educational needs directly into training. The issues involved have been discussed at length in MDRC reports (cf. Gueron

and Pauly, 1991) and are well understood. The problem is that the research design adopted for JOBS includes investigation of these two alternatives at only three sites: Fulton County, Georgia; Kent County, Michigan; and Riverside County in California (MDRC, 1992). There is reason to believe that the relative productivity of the two tracks will be conditioned not only by client characteristics, but by program and economic environment factors as well. As a result, it would seem that the value added from investigating this fundamental issue at more than three sites might be substantial.

Sanctions. Available evaluations show substantial differences across welfare-to-work demonstrations in the incidence of sanctions applied to clients who fail to comply with program rules. At the same time, there is some indication of association between a high incidence of sanctioning and program impacts, most notably in demonstrations in Riverside and San Diego Counties in California.[21] Some policy analysts, most notably Larry Mead, have argued that rigorous administration of employment preparation requirements is an essential component of effective WRET policy, and willingness to sanction is an important manifestation of such rigor (Mead, 1992). As Greenberg and Wiseman (1992b) have pointed out, however, the literature on sanctioning typically confuses outcomes with inputs. The incidence of sanctioning is an outcome that is the product of the interaction of a sanctioning policy with an environment that consists of economic circumstances, client characteristics, and particular program structure. We simply do not know what the net effects of sanctioning policy are, but there is general agreement that the issue is important.

Sanctioning policy is a good example of a program option that is not easily tested at a single site. This is because the very nature of the policy is communicated and reinforced as much through community understanding as it is through one-to-one interaction with particular clients. As a result, random assignment of clients at a single site to rigorous versus lax sanctioning regimes cannot be expected to provide results that can be used to identify what would occur if rigorous enforcement were an agency-wide activity.

Labor Market Intermediation. Subcontracting is a common feature of most current welfare-to-work programs, but in general the subcontracts involve public and private nonprofit agencies that provide training

programs. It is far less common to contract for labor market intermediation, that is, the placement of recipients in employment and support of those clients during the early phases of employment adjustment. Over the past five years, a number of private (and in some cases, for-profit) firms have claimed to achieve substantial results by playing an aggressive role in labor market intermediation. These accomplishments are of interest for at least two reasons. First, there is a presumptive case that assistance is needed in accomplishing the myriad tasks attendant to moving some workers into full-time employment. Second, labor market intermediation is a natural complement to the normal activities of those businesses that provide temporary employment services. The temporary employment services industry has exceptional knowledge of, and access to, local employers, and at least anecdotal evidence suggests that many temporary placements eventually become permanent hires. Because of the importance of long-term service relationships to the profitability of such firms, there is considerable incentive to make sure that persons placed do indeed perform well. It makes sense to investigate this programmatic option in a systematic way. Given that program models already exist and implementation has been demonstrated to be feasible, the next concern would seem to be to assess impacts in the context of a greater variety of administrative and economic environments.

Note that this innovation focuses on a mode (call it "privatization") of delivery, not on the technique pursued by the labor market intermediary, and the results will again be something of a black box, although presumably the interaction of black box with client characteristics and economic environment will be better understood. Nonetheless, the intervention seems to meet most of the requirements set out earlier.

Space does not permit a full working out of any of these alternatives, but it is worth noting an example of a possibility that is not included, and why this exclusion occurs. One element of the current GAIN evaluation that is not included in our list of candidates for multisite evaluation is experimentation with alternative case management techniques. Broadly speaking, the choice here is between case management structures that emphasize continuous oversight of client activity by a single caseworker and management structures that assure a sequence of activities but devote fewer resources to frequent interaction with clients

(Doolittle and Riccio, 1992). It appears to us that the gains from intensive case management, if present, are probably principally influenced by client characteristics. At the same time, achieving intensive case management is a difficult problem for which well-documented models are in short supply. In this context, it makes sense to concentrate on treatment implementation in a few sites with sufficient variation in client characteristics to determine whether variation in treatment along this dimension has any consequence at all and, if so, for whom. Generalization to multiple sites would appear to produce little payoff at the cost of high risk of failure to implement the necessary intervention.

Conducting Multisite Evaluations: The Hard Facts

The agenda proposed above has been selected in part given consideration of the likelihood that sites could be recruited for the effort. But no matter how skillful the diplomatic capabilities of the sponsoring federal agencies, successful multisite evaluation of treatment variations of the sort we propose would ultimately require some imposition of conditions on the state and local agencies that administer WRET programs. In part, these conditions are intended to keep the number of sites as small as possible and are implied by the model developed in section 3. They are also intended to address practical problems encountered by the recent multisite WRET demonstrations described in section 4. No matter how readily justified, the conditions can be substantial. Examples of the sorts of conditions that might arise in the context of a multisite demonstration appear below:

(1) Sites would be randomly selected to participate in evaluations, with the probability of selection related to site size. The right of individual sites to refuse to participate would be limited.

(2) Selected sites would be randomly assigned a specified "treatment." The right of individual sites to vary from implementing the assigned treatment in the manner specified would be constrained.

(3) Unplanned variations--for example, in the size of training classes--and adaptive adjustments by sites would be limited.

(4) Similar data on program inputs and outcomes would be required from each site.

(5) Each site would be required to implement the evaluation design (for example, random assignment) in the manner specified.

These conditions are obviously highly restrictive. In the welfare-to-work demonstrations conducted during the 1980s and in those currently being conducted, states and local welfare agencies have essentially selected themselves for evaluation and determined the treatment to be evaluated (Wiseman, 1993). Often they even determined the rigor with which they were evaluated (for example, whether random assignment was used.) Clearly the willingness of the federal government to sanction a wide range of interventions encouraged innovation, letting, to use Chairman Mao's term, "a thousand flowers bloom."

The problem is that the flowers that are produced by unconstrained state innovation don't make much of a bouquet, and they wilt. Because without direction state demonstrations vary so substantially, multilevel evaluation appears unlikely to succeed under such a voluntary regime. Instead, a necessary condition for getting into the black box is that the federal government assume greater responsibility for imposing the necessary constraints, perhaps through a combination of positive incentives and the threat of fiscal sanctions. Moreover, federal resources would be required to monitor compliance and to provide necessary technical assistance to state and local welfare agencies. Such steps may, of course, may be impractical or politically infeasible. But without them it seems unlikely that much progress can be made in refining the empirical basis for welfare-to-work policymaking.

Proposals for multisite evaluations are likely to be resisted in part because of the difficulties encountered in implementing the JTPA evaluation already described. But many of the reasons for site nonparticipation reported by Hotz (1992, p. 95) reflect concern about the effect of the experience on agency compliance with other federal and state performance standards. Surely if such standards conflict with implementation of a federally sponsored demonstration, they can be waived. It should also be pointed out

that while the demonstrations we propose require more sites, they do not require the same operational scale as used in the JOBS evaluations.

6. CONCLUSIONS

In 1992, the nation's research strategy for the support of welfare policy had two components. On the one hand, the Manpower Demonstration Research Corporation was coordinating a six-site study of the effects of JOBS with an average of 8,000 people per site. On the other, the Bush administration was encouraging states to propose their own innovations with virtually no restriction on content. The JOBS study promised more evidence on site mean effects, and some improvement in treatment specification, but, insofar as the research plan has been reported to date, little improvement in the specification and measurement of treatments or in understanding of the variable interactions that produce the site effects. As a result, while the JOBS evaluation may satisfy the congressional mandate to ascertain the net impact of the JOBS intervention, we are uncertain about its productivity in enhancing our understanding of how such programs work or how they are best managed. Given the very limited comparability among the other state demonstrations, little additional information can be expected from them.

Faced with this situation, welfare policymakers may opt, as others have done in the past, for pushing some sort of comprehensive welfare reform scheme that will render the kind of evaluations discussed in section 4 largely irrelevant. But past experience also shows that, whatever is pursued as welfare reform, certain basic issues will resurface. Given that it is likely that any new reform scheme will still incorporate efforts at encouraging self-support through employment and training, it is likely that, sooner or later, interest in research in this area will be rekindled. Since 1981, research on welfare-to-work initiatives has been largely conducted in response to state initiatives. A more proactive policy would attempt to better organize research and to create the necessary incentives to assure that the research is carried out. This paper has attempted to create the general statistical framework for thinking about how this should be done.

There are several obvious next steps.

(1) There is still much work to do on our evaluation model. In particular, we must address the small-sample properties of the statistical test we have applied for evaluating the homogeneity of site effects, and we need to develop more formally the consequence of lack of precision in measurement of inputs.

(2) The results we have developed for evaluating the required number of sites for macro-equation estimation needs more work, and we think it will be useful to apply the macro equation to an evaluation of the gains from adding (or, for that matter, deleting) an additional site to a demonstration set. Several research efforts in progress might benefit from this type of analysis, for example the multisite evaluation of the effects of changing the definition of unemployment used to determine eligibility for the two-parent component of Aid to Families with Dependent Children.[22]

(3) Effort is needed to develop input measures for the services that commonly make up welfare-to-work programs. We have argued above that better measures will enhance replicability and allow greater precision of impact estimation for a given number of sites or, alternatively, achievement of the same precision with fewer sites.

(4) As discussed above, implementation of research designs is often hampered by the absence of ground-level incentives for agencies to participate and agency staff to comply with operations requirements. Mandating such compliance is often difficult, and attention needs to be given to the structuring of incentives for proper operation of WRET experiments. MDRC's strategy is in part to emphasize assurance of test power at the site level. For example, the GAIN design assured that small effects could be detected with precision on a county level. Can site administrators be motivated by influences other than the lure of significance at the 5 percent level? We don't know. But at the same time, it is not obvious that site-level significance cannot be traded for the honor of collaboration in an experiment in which effects could possibly be found to be favorable and homoge-

neous *and* in which the size of the required effort was much smaller. In our "reasonable" case, Riverside county would be asked to contribute 75 treatment and control group members in total; in the GAIN demonstration 1,051 participants were in the control group alone.

(5) The "site" is an essential element of our model, but we have not formally explained what a site is. The key is that the term as used here refers to basic units of program administration.[23] These units should cover a geographic area small enough that a single set of environmental variables can be used to describe the program circumstances. To our knowledge, no catalog of such sites with summary characteristics data exits, but one would be invaluable as background for plotting research design alternatives, especially if attempts are made to assign techniques at random among sites and to design incentives appropriate to the nature of the unit of government responsible for program implementation in each.

Appendix A: The Optimum Number of Sites and Individuals per Site

Given (6), the equation for the precision of $\hat{\gamma}_1$, and (11), the cost function, it is possible to solve for the optimum number of sites (J^*) and the optimum number of individuals per site (n^*), given a target level of precision. As indicated in the text, the optimum number of sites is composed of two terms, J_1^* and $J_2(n^*)$. The first term does not depend on n^* and thus can be computed directly. The optimum value of J_1 is given in the text by (8).

In order to solve for n^* and $J_2(n^*)$, it is useful to rewrite (7) as

$$J^* = J_1^* + \frac{Q}{(n - K_1)}, \tag{A1}$$

where

$$Q = \frac{\sigma_e^2/s_P^2}{\sigma_1^2 s_{FI}^2 (1 - \overline{R}_1^2)} . \tag{A2}$$

Substituting (A1) into the cost function yields a cost function that depends on *n* and fixed parameters (including J_1^*):

$$C = k(m + n)\left(J_1^* + \frac{Q}{(n - K_1)}\right). \tag{A3}$$

The first and second derivatives of the cost function are given by

$$\frac{\partial C}{\partial n} = k\left\{J_1^* - \frac{(m+n)Q}{(n-K_1)^2} + \frac{Q}{(n-K_1)}\right\} \tag{A4}$$

$$\frac{\partial^2 C}{\partial n^2} = \frac{2kQ}{(n-K_1)^3}\{m + K_1\}. \tag{A5}$$

Since the second derivative is always positive, the cost function is U-shaped with respect to *n*. Hence, the solution obtained by setting (A4) equal to zero and solving for *n* is the optimum (cost-minimizing) value of *n*, that is

$$n^* = \left[\frac{Q(m+K_1)}{J_1^*}\right]^{1/2} + K_1 \ . \tag{A6}$$

This expression simplifies to (12) if $K_1 = K_2 = 0$. Given the relative sizes of K_1, K_2, and n typical of most applications, the approximation provided by (12) provides a good approximation for (A6). For example, for the "reasonable" case hypothesized in Table 1, the exact value (rounded to a whole person) for n^* is 103; the approximation is 110.

The optimum value of J_2 is obtained by substituting n^* into (9). The optimum number of sites, in total, is given by $J^* = J_1^* + J_2^*$, and the total number of individuals across all sites is given by $N^* = n^* J^*$.

Notes

[1]Equation (1) and many of the equations that follow involve vector multiplication. For simplicity of notation, we have eliminated specific notational reference to the necessary vector transpositions.

[2]Like most advertised novelties, this one is more than a bit old hat, since what we are constructing can be interpreted as an "error components" model in which program outcomes reflect "within group (site)" and "across group" effects. See Amemiya (1978) and Hsiao (1986, pp. 151-153). (We thank Robert Moffitt for bringing this work to our attention.) More efficient estimates of the θ_j can be obtained by estimating equation (1) as is, rather than using the simple difference in site means overall or by subgroup. However, it is not essential to include X in the model if individuals are assigned randomly to the treatment and control groups, since random assignment implies that X and the participation indicator are uncorrelated. The primary advantage of the random assignment approach is that in principle it guarantees that the treatment and control groups are drawn from the same population (Burtless and Orr, 1986). The use of random assignment is a key element in the methodology employed by the Manpower Demonstration Research Corporation in its studies of welfare-to-work programs; see Gueron and Pauly (1991). Critical assessments of the role of random assignment in social program evaluation are presented in several papers in Manski and Garfinkel (1992).

[3]While the model developed here may be new to the WRET literature, similar analyses have appeared in the education literature, most notably in the work of Larry V. Hedges (1982a, 1982b). Stigler (1986; cited in Hedges, 1992) reports discovery of structurally similar multilevel analyses of multiple research studies in nineteenth-century astronomy. Rubin (1992) refers to the macro equation as the "effect-size surface."

[4]Similarly, a set of macro equations could be used to probe the determinants of β_j. We concentrate here on analysis of program effects.

[5]See Rubin (1992) for a similar proposition stated within the framework of meta-analysis.

[6]Equation (5) is written as it stands to highlight the consequences of multicollinearity, sample size, etc.

To see that the equation is correct, note that the variance of $\hat{\theta}_j$ is given by $\sigma^2_{e_j} / RSS_j$, where RSS_j is the residual sum of squares from an auxiliary regression of P_{ij} on X_{ij} for site *j* (Maddala, 1988, p. 101). Our equation follows automatically from the definition of corrected R^2, $\bar{R}^2 = 1 - [RSS_j / (n_j - K_1)] / S^2_{P_j}$.

[7]Note that $s^2_{Pj} = s_j(1 - s_j)$, where s_j is the fraction of individuals in the participant group and $(1 - s_j)$ is the fraction of individuals in the control group. Hence, $s^2_{Pj} = 0.25$ if the participant and the control groups are the same size.

[8]This exercise may be compared to the pioneering work of Conlisk and Watts (CW) on the experimental design for the New Jersey Income Maintenance Experiment (Conlisk and Watts, 1969). The New Jersey experiment had but one general geographic location, but the household sample drawn from this site was stratified on the basis of prior income, and households selected for experimental treatment were assigned to one of a number of combinations of the negative income tax guarantee and the rate at which benefits declined as household income increased. Recast in terms of multilevel modeling, these different combinations (called "design points" by CW) can be thought of as "sites," and the CW optimization problem involved choosing sample sizes across the sites in such a way that efficiency of the estimates of the various macro parameters for the problem--the income and price effects of the negative tax on labor supply--was maximized given a fixed demonstration-wide budget constraint. In the CW case costs varied substantially across "sites" because of the variation in the income guarantee and tax rate; in our model costs are the same in all sites, and the question is how many "design points" to have. The CW model implicitly assumes fixed effects of the given program parameter combinations, while the multilevel model combines (mixes, in common terminology) fixed (the F_j) and random effects (the w_j). See Hedges (1992).

[9]See Appendix A. The approximation is quite accurate if *n* and *J* are large relative to K_1 and K_2, respectively.

[10]The source of this figure is discussed later in the paper: see footnote 18.

[11]Moffitt (1992) also reports similar site requirements for evaluating the consequences of WRET demonstrations for the likelihood that persons will apply for assistance.

[12]The material that follows is adapted from Hedges (1984), pp. 34-35. We are grateful to Daniel Friedlander of the Manpower Demonstration Research Corporation for providing the GAIN data upon which the computations that follow are based.

[13]We discuss the problem of characterization of program inputs below.

[14]See Manpower Demonstration Research Corporation (1992). The novelty of the JOBS evaluation is inclusion of both "net" and "differential" impact studies. The sites for which net impact evaluation is to be done will feature random assignment between control and treatment groups. Those in which differential impact estimates are to be obtained have, in addition to the control group, two treatment groups differentiated by either the emphasis placed on up-front job search or the organization of case management.

[15]H_T values for adult men and women are 10.9 and 10.3 respectively, with the critical χ^2 value of 25. For out-of-school female and male youths H_T is 10.6 and 15.1, with a critical χ^2 value of 23.7. We are grateful to Larry Orr and Winston Lin of Abt Associates, Inc., for providing the data for these calculations.

[16]Nonetheless, an attempt to estimate such equations has been made, but the report has not been released (it is forthcoming).

[17]We are grateful to Fred Doolittle of the Manpower Demonstration Research Corporation for information on JTPA study implementation problems.

[18]Cost estimates for Abt Associates were provided by Larry Orr; costs for MDRC operations and data on site subsidies provided by the U.S. Department of Labor were provided by Fred Doolittle. We divided DOL costs between fixed and variable components by regressing the site payments on a constant term and the observation count, and a term representing the number of organizations at each site that were involved in random assignment. $\bar{R}^2$ for the equation was .94.

[19]Variation in programmatic inputs will largely reflect site budgetary allocation decisions. For example, one site may decide to allocate most of its training and employment budget to relatively low cost services that are received by a large number of program participants, while another site with an identical budget emphasizes high-cost services that are received by relatively few persons (Friedlander and Gueron, 1992). Even more fundamentally, sites must determine the share of their budgets to allocate to service provision and the share to allocate to administrative tasks. Among the latter are assessment of employability and educational abilities, case management, and in the case of mandatory programs for welfare recipients, compliance monitoring.

[20]"Commensurate" means here that the return on an investment in knowledge acquisition of this type is comparable to that expected from other uses of government funds. For welfare-related employment and training programs, such returns are often hard to assess, and there are to our knowledge no examples in the literature of attempts to perform such evaluations. Allan S. Detsky (1989, 1990) has published a number of studies of the cost-effectiveness of clinical trials in medicine.

[21]See Hamilton and Friedlander (1989) and Riccio and Friedlander (1992). The association of Riverside County's substantial GAIN impacts and its sanctioning policy has created considerable controversy.

[22]For a description, see Wiseman (1993). In principle a multisite demonstration could be used to study program effects on the rate at which persons not receiving public assistance go on welfare (Moffitt, 1992). However, an investigation of such entry effects would not involve random assignment within sites at all but rather random assignment of program types across a sample of all sites. Demonstrations with random assignment of participants within sites have uncertain consequences for entry because they essentially create a lottery, and we have no idea how such programs are perceived by potential participants.

[23]For example, at a local level, the JTPA program is operated by over 600 service delivery areas (SDAs) and the AFDC and Food Stamp programs are administered by around 1,600 local agencies.

References

Amemiya, Takeshi. 1978. "A Note on a Random Coefficients Model." *International Economic Review*, 19(3), 793-796.

Auspos, Patricia, and Kay E. Sherwood. 1992. *Assessing JOBS Participants: Issues and Trade-Offs*. New York: Manpower Demonstration Research Corporation.

Blank, Rebecca. 1994. "The Employment Strategy: Public Policies to Increase Work and Earnings," in Sheldon H. Danziger, Gary D. Sandefur, and Daniel H. Weinberg, editors, *Poverty and Public Policy: What Do We Know? What Should We Do?* Cambridge, Massachusetts: Harvard University Press, forthcoming.

Bloom, Howard S. 1991. *The National JTPA Study: Baseline Characteristics of the Experimental Sample*. Bethesda, Maryland: Abt Associates, Inc.

Bloom, Howard S., Larry Orr, George Cave, Stephen H. Bell, and Fred Doolittle. 1992. *The National JTPA Study: Title II-A Impacts on Earnings and Employment at 18 Months, Executive Summary*. Bethesda, Maryland: Abt Associates, Inc.

Bryk, Anthony S., and Stephen W. Raudenbush. 1992. *Hierarchical Linear Models*. Newbury Park, California: Sage Publications.

Burtless, Gary, and Larry Orr. 1986. "Are Classical Experiments Needed for Manpower Policy?" *The Journal of Human Resources*, 21(4), 606-639.

Conlisk, John, and Mordecai Kurz. 1972. *The Assignment Model of the Seattle and Denver Income Maintenance Experiements*. Menlo Park, California: Stanford Research Institute, Research Memorandum 15.

Conlisk, John, and Harold Watts. 1969. "A Model for Optimizing Experimental Designs for Estimating Response Surfaces." *American Statistical Association Proceedings, Social Statistics Section*, 150-156.

Cook, Thomas D., and Harris Cooper, David S. Cordray, Heidi Hartmann, Larry V. Hedges, Richard J. Light, Thomas A. Louis, and Frederick Mosteller. 1992. *Meta-Analysis for Explanation: A Casebook*. New York: Russell Sage Foundation.

Detsky, Allan S. 1989. "Are Clinical Trials a Cost-Effective Investment? *Journal of the American Medical Association*, 262(13), 1795-1800.

_____. 1990. "Using Cost-Effectiveness Analysis to Improve the Efficiency of Allocating Funds to Clinical Trials." *Statistics in Medicine*, 9, 173-184.

Doolittle, Fred, and James Riccio. 1992. "Case Management in Welfare Employment Programs," in Charles F. Manski and Irwin Garfinkel, editors, *Evaluating Welfare and Training Programs*. Cambridge, Massachusetts: Harvard University Press, 1992, 310-343.

Friedlander, Daniel, and Judith M. Gueron. 1992. "Are High-Cost Services More Effective than Low-Cost Services?" in Charles F. Manski and Irwin Garfinkel, editors, *Evaluating Welfare and Training Programs*. Cambridge, Massachusetts: Harvard University Press, 1992, 143-198.

Garner, Catherine L., and Stephen W. Raudenbush. 1991. "Neighborhood Effects on Educational Attainment: A Multilevel Analysis." *Sociology of Education*, 64(3), 251-262.

Greenberg, David, and Michael Wiseman. 1990. *Work-Welfare Initiatives: How Do We Add Things Up?* Unpublished paper presented at the Twelfth Annual Research Conference of the Association for Public Policy and Management.

_____. 1992a. "What Did the OBRA Demonstrations Do?" in Charles F. Manski and Irwin Garfinkel, editors, *Evaluating Welfare and Training Programs*. Cambridge, Massachusetts: Harvard University Press, 1992, 25-75.

_____. 1992b. "What Did the Work-Welfare Demonstrations Do?" Institute for Research on Poverty Discussion Paper #969-92, University of Wisconsin-Madison.

Gueron, Judith M., and Edward Pauly. 1991. *From Welfare to Work*. New York: Russell Sage Foundation.

Hamilton, Gayle, and Daniel Friedlander. 1989. *Final Report on the Saturation Work Initiative Model in San Diego*. New York: Manpower Demonstration Research Corporation.

Hedges, L. V. 1982a. "Estimation of Effect Size from a Series of Independent Experiments." *Psychological Bulletin*, 92, 490-499.

_____. 1982b. "Fitting Continuous Models to Effect Size Data." *Journal of Educational Statistics*, 7, 245-270.

_____. 1984. "Advances in Statistical Methods for Meta-Analysis," in William H. Yeats and Paul M. Wortman, editors, *Issues in Data Synthesis*. San Francisco: Jossey-Bass, 1984, 25-42.

_____. 1992. "Meta-Analysis." *Journal of Educational Statistics*, 17(4), 279-296.

Hedges, L. V., and I. Olkin. 1985. *Statistical Methods for Meta-Analysis*. New York: Academic Press.

Hotz, V. Joseph. 1992. "Designing an Evaluation of the Job Training Partnership Act," in Charles F. Manski and Irwin Garfinkel, editors, *Evaluating Welfare and Training Programs*. Cambridge, Massachusetts: Harvard University Press, 1992, 76-114.

Hsiao, Cheng. 1986. *Analysis of Panel Data*. New York: Cambridge University Press.

Hunter, John E., and Frank L. Schmidt. 1990. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Newbury Park, California: Sage Publications.

Kirk, Roger E. 1982. *Experimental Design: Procedures for the Behavioral Sciences*, second ed. Pacific Grove, California: Brooks/Cole Publishing Co.

Levitan, Sar A. 1992. "The Evaluation Industry Impacts Policy." *Workforce*, 1(3), 34-45.

Maddala, G. S. 1988. *Introduction to Econometrics*. New York: Macmillan Publishing Co.

Manpower Demonstration Research Corporation. 1992. *The JOBS Evaluation: Overview of the Design*. New York: Manpower Demonstration Research Corporation.

Manski, Charles F., and Irwin Garfinkel. 1992. *Evaluating Welfare and Training Programs*. Cambridge, Massachusetts: Harvard University Press.

Moffitt, Robert. 1992. "Evaluation Methods for Program Entry Effects," in Charles F. Manski and Irwin Garfinkel, editors, *Evaluating Welfare and Training Programs*. Cambridge, Massachusetts: Harvard University Press, 1992, 231-252.

Puma, Michael J., Nancy R. Burstein, Katy Merrell, and Gary Silverstein. 1990. *Evaluation of the Food Stamp Employment and Training Program: Final Report*. Bethesda, Maryland: Abt Associates, Inc.

Riccio, James, and Daniel Friedlander. 1992. *GAIN: Program Strategies, Participation Patterns, and First-Year Impacts in Six Counties*. New York: Manpower Demonstration Research Corporation.

Rosenthal, Robert. 1991. *Meta-Analytic Procedures for Social Research*, revised edition. Newbury Park, California: Sage Publications.

Rosenthal, Robert and D. B. Rubin. 1982. "Comparing Effect Sizes of Independent Studies." *Psychological Bulletin*, 92, 500-504.

Rubin, Donald B. 1992. Meta-Analysis: Literature Synthesis or Effect-Size Surface Estimation? *Journal of Educational Statistics*, 17(4), 363-374.

Stigler, S. M. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, Massachusetts: Harvard University Press.

Wiseman, Michael. 1993. "The New State Welfare Initiatives." Institute for Research on Poverty Discussion Paper, University of Wisconsin-Madison (in press).